基于概念基元的词语相似度计算研究

池哲洁*①② 张 全②
①(中国科学院大学 北京 100049)
②(中国科学院声学研究所 北京 100190)

摘 要：词语相似度的计算在机器翻译、信息检索等多个领域有重要作用。该文以概念层次网络理论的概念基元符号系统为语义资源，在共性与差异性对比思想下，提出一个涵盖层次性、网络性、对比对偶特性、挂靠特性及五元组信息的多维度词语相似度计算方法；在节点深度和节点距离度量上，引入权重以增加不同层次间的区分程度。在人工打分的测试集上进行实验，结果表明该方法计算的相似度与人工判断的符合程度较好，兼容度、相关系数和序对符合度分别达到 0.812, 0.786 和 0.775；同时，相关性检验的结果也显示该方法的计算值与人工打分显著相关。

关键词：词语相似度；语义距离；概念层次网络；概念基元
中图分类号： TP391 **文献标识码**： A **文章编号**：1009-5896(2017)01-0150-09
DOI: 10.11999/JEIT160176

Word Similarity Measurement Based on Concept Primitive

CHI Zhejie①② ZHANG Quan②
①(*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
②(*Institute of Acoustics*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: Word similarity measurement plays an important role in machine learning, information retrieval and many other fields. Regarding the concept primitive symbol system of Hierarchical network of concepts theory as semantic resource and comparing commonness with difference, a multi-dimensional computational method for similarity is proposed which considers the hierarchy, netted nature, comparability and duality, attached feature and quintuple information of the system. Weight strategy is introduced for node depth and distance measurement to increase the discrimination of node level. Experiment on manual scoring test set shows that the computed similarities are consistent with human judgments. The proposed method achieves 0.812, 0.786, and 0.775 in compatibility degree, correlation coefficient, and ordinal pair conformity respectively. Meanwhile, the result of correlation test further proofs that the computed similarities and human's scores are significantly correlated.

Key words: Word similarity; Semantic distance; Hierarchical network of concepts; Concept primitive

1 引言

词语相似度的计算在机器翻译、信息检索、自然语言处理等多个领域具有重要作用。相似度反映两个事物间特征的重合程度，而词语涉及多方面特征，包括词法、句法、语义及语用等，但语义在相似度中的影响最大，因此，本文考虑词语语义相似度，主要是指词语在语义概念上的重合程度。在度量上，Lin[1]认为两个词语的相似度取决于它们的共性和差异性，并以信息论角度提出了相似度的计算方法： $S(A,B)=I(\text{common}(A,B))/I(\text{description}(A,B))$，其中，$I(\text{common}(A,B))$ 为词语间的共性信息，$I(\text{description}(A,B))$ 为词语间的描述性信息，一般由共性和差异性组合而成。Lin 的方法给出了重合度度量的一个通用思想，本文以此为基础设计相似度的计算方法，并将相似度值限制在[0,1]。

目前，进行词语语义相似度计算主要有两种方法，一种是利用依托某种世界知识所构建的语义词典的方法，另一种则是基于大规模语料统计的方法。基于语义词典的方法主要利用语义词典将词汇按照语义类别组织在树状层次结构中的特点，考虑其中概念节点间的上下位或同位关系等，通过距离或信息内容来度量词语间的相似性。英语词语相似度计算主要基于 WordNet，文献[2]通过考虑概念词与其

收稿日期：2016-02-25；改回日期：2016-09-14；网络出版：2016-11-14
*通信作者：池哲洁 chizhejie@sina.com
基金项目：国家 863 计划“十二五”项目(2012AA011102)，国家语委“十二五”科研项目(YB125-53)
Foundation Items: The Twelfth Five-Year Project of National 863 Program of China (2012AA011102), The State Language Commission Twelfth Five-Year Research Project (YB125-53)

最近公共父节点概念词的位置关系来计算相似度；Resnik[3]提出了直接利用最大公共祖先节点概念词的信息内容来计算相似度的方法；其后的改进方法一般都额外考虑节点深度、密度、语义重合度、概念频数、语义数目等因素进行综合计算[4,5]。汉语词语相似度计算多是采用《知网》来开展的，刘群等人[6]探索《知网》义原体系，采用上下位关系度量义原间的距离，利用距离和相似度成反比例的关系设计义原相似度计算公式，然后将词语整体相似度分解成多个义原对相似度的组合，对部分义原相似度进行加权平均得到词语的整体相似度；李国佳[7]采用义原信息量来计算概念间主类义原的相似度，并结合义原角色关系综合计算词语相似度；张沪寅等人[8]通过义原距离限制义原深度对相似度的影响而实现义原相似度计算的改进；孙晶等人[9]提出逆概念频率计算方法，并用于为不同义原定义权重，根据动态权重计算词的相似度。基于大规模语料统计的方法是建立在相似词语所处的上下文环境是相似的假设上，将词语相似度的计算转移到它们所处上下文环境的对比中来。Brown 等人[10]基于平均互信息计算词语相似度；关毅等人[11]提取词语的上下文概念分布信息，利用相关熵进行差异比较，从而计算语义相似度；王石等人[12]采用词汇在二词短语中的搭配词作为其上下文，在自动构建大规模二词短语的基础上，使用 tf-idf 作为向量权重，构造直接和间接搭配向量，通过计算搭配向量间的夹角余弦，将其作为词语相似度结果。上述两种方法各有优缺点，基于语义词典的方法简单有效、直观且易于理解，但需要有完备的知识库支撑，一般人为构建的语义词典具有一定局限性；另外，它对于不包含在词典中的词语(未登录词)基本不具备处理能力。基于语料库统计的方法比较客观，并且没有未登录词处理的问题；不过其对训练语料的依赖性大，理论上，所使用的语料库对真实语言的代表性越好，则计算结果与实际越符合，但在实践中，构建这种“完美代表性”的语料库难度巨大；另外，该方法一般计算量较大，容易受数据稀疏和数据噪声的影响而出现错误。从优缺点角度出发，基于混合技术的词语相似度算法存在很大的发展空间[13]。

概念层次网络[14,15](Hierarchical Network of Concepts, HNC)理论是面向整个语言理解的理论框架，是中文信息处理的 3 个流派之一[16]。该理论立足于语言概念空间，通过构建概念联想脉络实现语言的理解。语言概念空间的基层是概念基元空间，其中包含一套概念基元符号系统，该符号系统由概念基元组成，具有层次性且采用基元化的语义定义，能够准确表达词语的语义内涵，适合当作语义词典使用。已有工作中，史燕[17]考虑了概念基元的层次性和五元组信息，基于距离计算概念基元相似度并利用组合符号实现词语相似度的计算；吴佐衍等人[16]根据 HNC 符号的编码规则和符号映射理论，综合考虑概念内涵、概念外部特征和概念类别信息，提出概念表示的相似度加权计算公式，然后考虑组合符号，实现 HNC 符号的相似度计算，最终利用词语和 HNC 符号的映射关系实现词语相似度的计算。上述工作中都考虑了概念基元符号系统的层次性以及一些外部特性，最终对词语相似度计算也起到了一定作用，但这些考虑并不全面，忽略了概念节点间的对偶、对比特性及网络性等对语义表示起很大作用的因素。本文以概念基元符号系统为基础，充分考虑语义网络设计的层次性及一些外部特性，同时考虑它们的对偶、对比特性及网络性，力图实现一个更加合理的词语相似度计算方法。

2　概念基元符号系统

本文的工作是以概念基元符号系统为基础，这里对该系统做基本介绍，更详细信息请参阅文献[14, 15]。

HNC 理论注重抽象概念的表达，对于具体概念，主张采用近似的方案进行描述。语义网络是为描述抽象概念而设计的，它是由符号化的概念基元构成。语义网络中的概念基元具有层次性，可以用树状结构进行组织，并将所有概念基元的根记为 LCS(Language Concept Space)。纵观整个语义网络，它是按“概念范畴－概念林－概念树－延伸概念结构”的方式从高到低进行组织的。以图 1 中的部分语义网络为例，a 处在概念范畴层，其下有多个子节点，a1 为其中一个，处于概念林层面，a14 是 a1 下的一个概念树节点，其下又包含一系列延伸概念节点。HNC 将概念树以上(包括概念树)部分称为概念基元的高层，在高层之后则是进入延伸结构，共两类，第 1 类延伸结构有 3 种：对偶性、对比性和包含性，第 2 类延伸结构也是 3 种：交织性延伸、并列性延伸和定向性延伸。

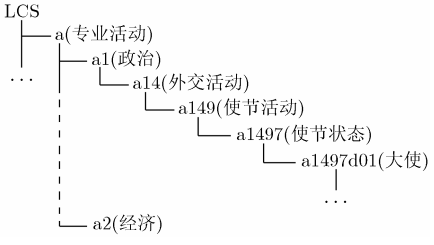

图 1 语义网络节点示例

概念基元表达了概念的内涵，通常情况下，同一个内涵会有多方面的表现，这在自然语言中表现为词性现象，而在 HNC 中则是通过五元组来体现的。五元组包括动态、静态、属性、值和效应，分别对应符号 v, g, u, z 和 r，将五元组符号作用到语义网络中的节点，则可形成同一内涵概念不同侧面的表示。

将语义网络中表示内涵的概念基元符号和表示外在特征的五元组符号基于一定规则组合起来形成 HNC 符号。由 HNC 符号表示概念的这一套系统我们称为概念基元符号系统。自然语言中的词语与概念基元符号系统中的符号存在映射关系，由人工将这类映射关系组织起来形成词语-概念基元映射表，使用该映射表，可以将词语相似度的计算转移到概念基元空间中。

语义网络中的概念基元除了具有层次性，还具有网络性，也就是关联性。一方面，上下位的层次关系本身带有关联信息；另一方面，处于不同子网络下的概念基元间也可能存在关联，如“a63e219”(综合分析性理论)与“811”(思维活动的综合与分析)有很大的关联性。这类不同网络间的关联无法通过表层符号直接体现出来，目前的一个解决方案就是人工构建这种隐式的关联，以概念关联式的形式呈现。HNC 理论定义了 10 种沿袭逻辑关联，此处将其总结如下(括号内为对应关联符号)：强关联(≡)、强交式关联(=)、强流式关联(<=)、强源式关联(=>)、包含(%=)、属于(=%)、对应(:=)、等同(=:)、定义(::=)、虚设(==)。

3 词语相似度计算方法

利用概念基元符号系统的组织和编码形式以及词语和 HNC 符号的映射关系，本文提出基于概念基元的词语相似度计算方法，主要步骤为：首先提出不含组合符号的 HNC 符号(称为单一 HNC 符号)间的相似度计算方法，进而考虑其组合形式的计算，最后利用词语和 HNC 符号的映射关系，实现词语的相似度计算。

3.1 单一 HNC 符号相似度计算

在 HNC 符号相似度计算上，本文将从层次性、对比对偶特性、网络性、挂靠特性及五元组信息等多个维度充分考虑各符号间的共性和差异性，在此基础上设计相似度计算公式。

为方便后文表述，这里先对一些名称的标记进行约定：用 hs 标记 HNC 符号，概念基元符号则记为 cp，五元组符号记为 ft，概念基元 cp_1，cp_2 的最大公共节点记为 $\mathrm{c}(\mathrm{cp}_1,\mathrm{cp}_2)$。另外，给出以下两个定义：

定义 1 节点深度：指语义网络中概念节点 cp 的最大层次数，记为 $\mathrm{de}(\mathrm{cp})$，而 $\mathrm{de}(\mathrm{LCS})=0$。

定义 2 节点距离：指连接两个概念节点 cp_1 和 cp_2 的最短路径长度，记为 $\mathrm{ds}(\mathrm{cp}_1,\mathrm{cp}_2)$。

3.1.1 概念基元相似度计算 基于共性和差异性对比的思想，概念基元间相似度计算的通用公式为

$$S(\mathrm{cp}_1,\mathrm{cp}_2)=\frac{\alpha\mathrm{cm}(\mathrm{cp}_1,\mathrm{cp}_2)+\beta}{\alpha\mathrm{cm}(\mathrm{cp}_1,\mathrm{cp}_2)+(1-\alpha)\mathrm{df}(\mathrm{cp}_1,\mathrm{cp}_2)+\beta} \tag{1}$$

其中，$\mathrm{cm}(\mathrm{cp}_1,\mathrm{cp}_2)$ 为 cp_1 和 cp_2 的共性度量表示，$\mathrm{df}(\mathrm{cp}_1,\mathrm{cp}_2)$ 则为其差异性度量表示，α 是调节因子，β 是平滑参数。从式(1)中可以看出，相似度通过计算共性在描述性中的占比获得，描述性信息则表现为共性和差异性之和，调节因子用于调整共性信息的权重，平滑参数的使用是为了调节共性度量值为 0 的情况。概念基元符号间存在共同的层次符号(至少根节点是一致的)，也存在有差别的部分，且这两部分都能够通过数值来度量，因此，基于共同的长度和相异的距离来度量相似性也是一种自然的想法。本文将最大公共节点的节点深度作为共性的度量，而差异性则通过节点距离来度量。另外，在节点深度及节点距离度量中，本文认为不同层次间的概念差异性是不一样的，一般越处底层其概念就越被细化，层次间差异性也越小。如图 1 所示，a1 到 a14 这一层的距离应该大于 a149 到 a1497 这一层。为体现层次间的这种差异，需要为节点的不同层次赋上权重(采用满足权重值与层次值成反比的函数进行拟合)。用 $\mathrm{de}_\mathrm{w}(\mathrm{c}(\mathrm{cp}_1,\mathrm{cp}_2))$ 表示 cp_1 和 cp_2 公共节点的带权节点深度，$\mathrm{ds}_\mathrm{w}(\mathrm{cp}_1,\mathrm{cp}_2)$ 表示带权节点距离，分别代入式(1)的 $\mathrm{cm}(\mathrm{cp}_1,\mathrm{cp}_2)$ 和 $\mathrm{df}(\mathrm{cp}_1,\mathrm{cp}_2)$ 部分，则可得到概念基元相似度计算的具体公式。

记节点层次的权重模拟函数为 $f(l)$，是层次值 l 的函数，概念基元 cp 的总层次数为 L，则 $\mathrm{de}_\mathrm{w}(\mathrm{cp})$ 的计算公式为

$$\mathrm{de}_\mathrm{w}(\mathrm{cp})=\sum_{l=0}^{L}f(l) \tag{2}$$

$\mathrm{ds}_\mathrm{w}(\mathrm{cp}_1,\mathrm{cp}_2)$ 则可通过式(3)方法计算：

$$\mathrm{ds}_\mathrm{w}(\mathrm{cp}_1,\mathrm{cp}_2)=\mathrm{de}_\mathrm{w}(\mathrm{cp}_1)+\mathrm{de}_\mathrm{w}(\mathrm{cp}_2)-2\mathrm{de}_\mathrm{w}(\mathrm{c}(\mathrm{cp}_1,\mathrm{cp}_2)) \tag{3}$$

3.1.2 对比、对偶特性的度量 语义网络中对相似度计算起主导作用的是层次性，不过其对比、对偶特性也不容忽视。对比性是指共寓于同一高层概念下的一组概念，彼此间存在量的差异；对偶性则是指一组概念彼此间存在质的差别。对比中处于两端的概念以及对偶中表示对立和对抗的概念常常构成反义关系，从而影响相似度；而对比其他位置的概念

及普通对称关系也会加大概念间的距离。本文将这两部分的作用考虑进概念间的差异性，通过适当放大概念间的距离来减小相似度。在计算上定义一个差异性缩放系数 w_{IM} 作用于带权节点距离 $\mathrm{ds_w}(\mathrm{cp_1},\mathrm{cp_2})$，$w_{\mathrm{IM}}$ 的计算方法如式(4)：

$$w_{\mathrm{IM}}=\begin{cases}\omega_{11}, & (\mathrm{cp_1},\mathrm{cp_2})\in A_1\\ \omega_{12}, & (\mathrm{cp_1},\mathrm{cp_2})\in A_2\\ 1.0, & \text{其它}\end{cases}\tag{4}$$

其中，$\omega_{11}>\omega_{12}>1.0$，集合 $A_1=\{(\mathrm{cp_1},\mathrm{cp_2})|\,\mathrm{cp_1},\mathrm{cp_2}$ 处于对比的两端或 $\mathrm{cp_1}$，$\mathrm{cp_2}$ 为对抗或对立概念$\}$，$A_2=\{(\mathrm{cp_1},\mathrm{cp_2})|\,\mathrm{cp_1}$，$\mathrm{cp_2}$ 为对比概念的但不处于两端或 $\mathrm{cp_1}$，$\mathrm{cp_2}$ 为普通对称概念$\}$。

3.1.3 网络语义关联的度量　除了层次性及对比、对偶特性，语义网络中还需要考虑的一个因素是网络性，即概念间的关联性质，此处主要考虑不同语义网络间的概念关联，这方面的度量需要借助已构建好的概念关联式。本文将关联性作用到概念间的共性 $\mathrm{de_w}(\mathrm{c}(\mathrm{cp_1},\mathrm{cp_2}))$ 上，采用一个共性缩放系数 w_{AT} 来度量不同的关联类型。本文考虑关联性较强的 8 种关联，将其分为 4 组，则 w_{AT} 的计算方法如式(5)：

$$w_{\mathrm{AT}}=\begin{cases}\omega_{21}, & \mathrm{at}\in B_1\\ \omega_{22}, & \mathrm{at}\in B_2\\ \omega_{23}, & \mathrm{at}\in B_3\\ \omega_{24}, & \mathrm{at}\in B_4\\ 1.0, & \text{其它}\end{cases}\tag{5}$$

其中，at 为概念关联符号，$\omega_{21}>\omega_{22}>\omega_{23}>\omega_{24}>1.0$，集合 $B_1=\{\equiv\}$，$B_2=\{=\}$，$B_3=\{<=,\ =>,\ \%=,\ =\%\}$，$B_4=\{:=,\ =:\}$。

3.1.4 挂靠类型语义关联的度量　挂靠是 HNC 概念表达的一种常用方式，在表示上，直接将一个概念符号与相关概念的符号拼接在一起。例如，表示“交通工具”的“pw22b”就是直接把具体概念“pw”(人造物)和基元概念“22b”(自身转移)连在一起。在挂靠表示中，向其他概念挂靠的概念称为挂靠层，而被挂靠概念称为本体层。挂靠层一般是表现概念的某些特性，其实质还是体现在本体层，因此，本文将挂靠层的区别放到差异性的考虑中，计算上采用一个差异性缩放系数 w_{AC} 来表示，考虑挂靠概念集合间的差异程度，w_{AC} 的计算公式为

$$w_{\mathrm{AC}}=\begin{cases}\omega_{31}, & C_1\cap C_2=\varnothing,\ C_1\cup C_2\neq\varnothing\\ \omega_{32}, & C_1\cap C_2\neq\varnothing,\ C_1\neq C_2\\ 1.0, & \text{其它}\end{cases}\tag{6}$$

其中，$\omega_{31}>\omega_{32}>1.0$，$C_1$ 为 $\mathrm{cp_1}$ 的挂靠集合，C_2 为 $\mathrm{cp_2}$ 的挂靠集合。

3.1.5 外在表现的度量　五元组是对概念不同侧面的表达，是概念外在表现的基元，在相似度计算上有具有一定影响。同一个概念基元作用上不同的五元组符号所表达的概念会有差别，本文将这部分的区别也反映到差异性中，设计一个差异性缩放系数 w_{FT} 来表示。依据五元组与语法学的词性大致对应关系(v 对应动词，u 对应形容词，g, z, r 对应名词)设置 w_{FT} 的计算式，表示为

$$w_{\mathrm{FT}}=\begin{cases}\omega_{41}, & (\mathrm{ft_1},\mathrm{ft_2})\in D_1\\ \omega_{42}, & (\mathrm{ft_1},\mathrm{ft_2})\in D_2\\ \omega_{43}, & (\mathrm{ft_1},\mathrm{ft_2})\in D_3\\ 1.0, & \text{其它}\end{cases}\tag{7}$$

其中，$\omega_{41}>\omega_{42}>\omega_{43}>1.0$，集合 $D_1=\{(\mathrm{ft_1},\mathrm{ft_2})|\,\mathrm{ft_1},\mathrm{ft_2}$ 对应词性全不同$\}$，$D_2=\{(\mathrm{ft_1},\mathrm{ft_2})|\,\mathrm{ft_1}$，$\mathrm{ft_2}$ 对应词性部分相同$\}$，$D_3=\{(\mathrm{ft_1},\mathrm{ft_2})|\,\mathrm{ft_1}$，$\mathrm{ft_2}$ 对应词性全相同但 g, z, r 有差别$\}$。

综合以上几个部分，可以得到 HNC 符号相似度的计算方法。在共性描述上，最终的计算公式为

$$\mathrm{cm}(\mathrm{hs_1},\mathrm{hs_2})=\begin{cases}\mathrm{de_w}(\mathrm{c}(\mathrm{cp_1},\mathrm{cp_2}))w_{\mathrm{AT}}, & \mathrm{de_w}(\mathrm{c}(\mathrm{cp_1},\mathrm{cp_2}))\neq 0\\ 0.5(w_{\mathrm{AT}}-1), & \mathrm{de_w}(\mathrm{c}(\mathrm{cp_1},\mathrm{cp_2}))=0\end{cases}\tag{8}$$

其中，cp_i 为 hs_i 中分解出的概念基元本体。当最大公共节点的带权节点长度为 0 时，为防止关联性不起作用，进行适当平滑。

差异性描述的最终计算公式为

$$\mathrm{df}(\mathrm{hs_1},\mathrm{hs_2})=\begin{cases}\mathrm{ds_w}(\mathrm{cp_1},\mathrm{cp_2})w_{\mathrm{IM}}w_{\mathrm{AC}}w_{\mathrm{FT}}, & \mathrm{ds_w}(\mathrm{cp_1},\mathrm{cp_2})\neq 0\\ 0.5(w_{\mathrm{IM}}+w_{\mathrm{AC}}+w_{\mathrm{FT}}-3), & \mathrm{ds_w}(\mathrm{cp_1},\mathrm{cp_2})=0\end{cases}\tag{9}$$

其中，平滑部分的考虑与共性描述的情况类似。沿用式(1)的形式，得到 HNC 符号的相似度计算公式为

$$S(\mathrm{hs_1},\mathrm{hs_2})=\frac{\alpha\mathrm{cm}(\mathrm{hs_1},\mathrm{hs_2})+\beta}{\alpha\mathrm{cm}(\mathrm{hs_1},\mathrm{hs_2})+(1-\alpha)\mathrm{df}(\mathrm{hs_1},\mathrm{hs_2})+\beta}\tag{10}$$

HNC 符号相似度的计算步骤如下(算法 1)：

(1)分解 HNC 符号，分别得到五元组符号、挂靠信息及概念基元信息；

(2)获取两个概念基元的公共节点，采用式(2)，式(3)计算公共节点带权节点深度及带权节点距离；

(3)分别考察概念间的对比、对偶特征，关联特性，挂靠信息及五元组信息，采用式(4)–式(7)计算 w_{IM}，w_{AT}，w_{AC} 及 w_{FT}；

(4)将以上各部分结果代入式(8)，式(9)求出共性和差异性描述信息，再利用式(10)计算最终结果。

3.2 组合结构分析及计算

前面提到的单一 HNC 符号只是概念海洋里的基本元素，更多的概念则是通过这些单一 HNC 符号组合进行表示的。HNC 定义了 12 种概念组合结构，分别是：作用(#)、效应($)、对象(&)、内容(|)、偏正(/)、主谓(‖)、展开(+)、并(,)、选(;)、一般逻辑组合(lyy)、非(!)、反(^)，各符号的具体含义请参见文献[15]。本文根据计算需要按组合符号所作用对象的数量将其分为两类：一元组合符号和二元组合符号，一元组合符号是指其作用对象只有一个，包括“非”和“反”，其余的则为二元组合符号。

不同的组合符号表示不同的意义，组合后的概念也各不相同，不过它们都会在一定程度上包含组合前概念的义项，因此，组合后的概念可以通过组合前的概念进行表示。本文在计算上先分解组合符号为单一符号，然后对组合符号赋予权重进行量化，最后采用加权求和的方法计算组合形式的相似度值。其中，一元组合符号计算如式(11)所示，二元组合符号计算公式为

$$S(\mathrm{hs}_1,\mathrm{hs}_2)=\lambda_i S(\mathrm{hs}_{11},\mathrm{hs}_2) \tag{11}$$

$$S(\mathrm{hs}_1,\mathrm{hs}_2)=\lambda_i S(\mathrm{hs}_{11},\mathrm{hs}_2)+(1-\lambda_i)S(\mathrm{hs}_{12},\mathrm{hs}_2) \tag{12}$$

其中，λ_i 为 hs_1 所包含的组合符号 ct 对应的权重，取值限定为 0 到 1；一元组合情况下，hs_1 分解出一个 HNC 符号 hs_{11}，二元组合情况下可分解成 hs_{11} 和 hs_{12}。另外，分解后的 HNC 符号本身可能还是一个组合形式，理论上，可以根据组合符号类型递归调用式(11)或式(12)，直到都分解成单一 HNC 符号求解的形式，不过本文在研究组合形式后，对此处计算做了如下修改(算法 2)：

(1)若两个待计算 HNC 符号相同，直接返回相似度值 1；

(2)若两个不同 HNC 符号分解后其组合类型相同，本文认为这种情况只需要分别考虑其对应部分的相似度，计算上则是先求出对应部分的相似度值后再加权求和，而不进行递归展开，如式(13)所示。

$$S(\mathrm{hs}_1,\mathrm{hs}_2)=\lambda_i S(\mathrm{hs}_{11},\mathrm{hs}_{21})+(1-\lambda_i)S(\mathrm{hs}_{12},\mathrm{hs}_{22}) \tag{13}$$

(3)其余情况则按式(11)或式(12)以递归形式展开计算。

3.3 词语相似度计算

实现 HNC 符号间的相似度计算并考虑组合结构情况后，就可以利用已构建好的词语和 HNC 符号映射表计算词语的相似度。计算时，先将词语映射到 HNC 符号，此时可能存在一个词语对应多个 HNC 符号义项的情况，本文规定相似度就取所有义项相似度的最大值，即

$$S(W_1,W_2)=\max_{i,j}\{S(\mathrm{hs1}_i,\mathrm{hs2}_j)\} \tag{14}$$

其中，W_1，W_2 是两个词语，其不同义项的 HNC 符号集合分别为 hs1 和 hs2，义项数分别为 m 和 n，即 $\mathrm{hs1}=\{\mathrm{hs1}_1,\mathrm{hs1}_2,\cdots,\mathrm{hs1}_m\}$, $\mathrm{hs2}=\{\mathrm{hs2}_1,\mathrm{hs2}_2,\cdots,\mathrm{hs2}_n\}$。

完整的词语相似度计算步骤如下(算法 3)：

(1)将 W_1 和 W_2 映射为 HNC 符号集 hs1 和 hs2 并分解出相应的义项集 $\{\mathrm{hs1}_i\}$ 和 $\{\mathrm{hs2}_j\}$；

(2)采用算法 2 循环计算两个集合中各项间的相似度；

(3)取步骤(2)中的最大值，作为最终相似度结果。

4 实验及分析

4.1 实验设置

目前对相似度计算进行评价的一个常见做法是将计算结果与人工评定的结果进行对比，通过定性及定量分析判断结果的好坏。在人工打分测试集选择上，本文使用文献[12]构建的测试集作为实验数据来源，该测试集仿照英语词语相似度基准测试集构建的方法，充分考虑了词语的“分布均匀性”和“相似均匀性”，具有较好的代表性。考虑到对比方法(将在后文介绍)使用的语义资源所收录的词语情况，本文从该测试集中剔除一些无法计算的词语对，最终保留 60 对词语用于测试(见表 1)。

在计算词语相似度前，需对相应参数进行设置，本文基于最优相关系数来设定参数，即选择使相关系数达到最优时对应的参数为最终的参数，表示为：$\boldsymbol{\Phi}_{\max}=\arg\max r(\boldsymbol{\Phi})$。实际计算中，首先结合经验为各参数设置取值范围，然后将取值范围空间按适当步长(本文使用的步长为 0.05)进行切分得到有限个离散点，最后在这些离散点中穷尽搜索最优参数，最终确定的各参数取值情况如表 2 所示(HNC 组合符号按#、$、&、|、/、‖、+、,、;、lyy、!、^顺序对应参数 $\lambda_1,\lambda_2,\cdots,\lambda_{12}$)。权重模拟函数取：$f(l)=1/\sqrt{l}\ (l\neq 0)$ 且 $f(0)=0$。另外，本文选取了 3 组方法作为对比实验，分别是文献[6]的方法(方法 1)，它是基于《知网》进行汉语词语相似度计算最为典型的方法，适合作为其他方法的比较标准；文献[12]的方法(方法 2)，该方法是基于大规模语料统计的新方法；文献[16]的方法(方法 3)，它是目前基于 HNC 理论考虑因素最全面、处理效果达到最好的方法；同时，将本文方法称为方法 4。

4.2 结果与分析

将 4 种算法分别应用于测试集，最终结果如表 1 所示，同时，将多组结果以折线图呈现(图 2)，

表 1 词语相似度计算结果

ID	W_1	W_2	S_{m}	S_1	S_2	S_3	S_4
1	汽车	轿车	0.8000	1.0000	0.3830	0.9239	0.9276
2	汽车	飞机	0.5230	0.4444	0.2050	0.5113	0.4688
3	汽车	医院	0.0150	0.1000	0.1950	0.4853	0.2523
4	珠宝	宝石	0.8920	0.1302	0.3670	0.4286	0.4991
5	珠宝	玻璃	0.4850	0.2015	0.1370	0.3909	0.4583
6	珠宝	正午	0.0150	0.0999	0.0070	0.2269	0.2943
7	中午	正午	0.9620	1.0000	0.3010	0.9435	0.9276
8	猫	狗	0.5850	1.0000	0.3710	0.8303	0.8108
9	猫	苹果	0.0460	0.2424	0.1290	0.6606	0.6324
10	香蕉	苹果	0.5690	1.0000	0.2550	0.7954	0.7728
11	男人	母亲	0.2620	0.8611	0.1650	0.3879	0.2485
12	男人	工作	0.1310	0.1197	0.0310	0.1658	0.1663
13	森林	林地	0.8190	0.1860	0.1770	0.8250	0.8483
14	森林	手机	0.0080	0.1860	0.1050	0.2050	0.1923
15	电话	手机	0.8040	1.0000	0.3700	0.8138	0.8818
16	电话	电视	0.4850	0.8960	0.1670	0.2158	0.4415
17	中国	联合国	0.4920	0.1364	0.1090	0.5114	0.5209
18	中国	公鸡	0.1150	0.1000	0.0350	0.3325	0.3147
19	绳	线	0.8230	1.0000	0.1690	0.3262	0.3116
20	绳	胳膊	0.0230	0.0882	0.0170	0.3216	0.2524
21	椅子	凳子	0.7690	1.0000	0.2980	1.0000	1.0000
22	椅子	风	0.0080	0.1263	0.0650	0.3244	0.1806
23	教师	科学家	0.5310	0.5759	0.2050	0.6606	0.5387
24	医院	诊所	0.8580	1.0000	0.1910	0.5250	0.8425
25	房子	桌子	0.3540	0.1534	0.1300	0.4658	0.2850
26	电影	邮票	0.0770	0.1667	0.0720	0.5587	0.2374
27	担心	忧虑	0.9810	1.0000	0.2330	0.8000	0.9449
28	担心	放心	0.4850	0.2857	0.1790	0.9500	0.6000
29	思考	考虑	0.9000	1.0000	0.2850	0.7303	0.7625
30	思考	问候	0.1230	0.1509	0.0000	0.3432	0.2661
31	跑	走	0.7540	0.6154	0.1630	0.7722	0.7251
32	跑	跳	0.6460	0.4444	0.2270	0.6516	0.8001
33	创造	发明	0.8960	0.6154	0.1320	0.4689	0.5835
34	创造	竞争	0.2000	0.2424	0.2800	0.3994	0.2703
35	告诉	通知	0.8150	1.0000	0.0280	0.5494	0.7812
36	告诉	衰老	0.0230	0.0964	0.0000	0.2575	0.2795
37	体会	感觉	0.8620	0.3478	0.2150	0.6990	0.7321
38	购买	销售	0.5310	0.2857	0.1620	0.5591	0.8168
39	停留	运动	0.3540	0.0444	0.0000	0.4542	0.4923
40	属于	存在	0.3380	0.1860	0.0630	0.3566	0.2008

表 1 词语相似度计算结果(续)

ID	W_1	W_2	S_{m}	S_1	S_2	S_3	S_4
41	高尚	崇高	0.9120	0.7884	0.1920	0.5849	0.6124
42	高尚	邪恶	0.3850	0.7884	0.0360	0.5455	0.6375
43	高尚	陡峭	0.0150	0.6160	0.0090	0.2841	0.2980
44	崎岖	陡峭	0.8190	0.6241	0.1050	0.6377	0.6311
45	崎岖	平坦	0.6080	0.8611	0.1190	0.5942	0.5948
46	聪明	机智	0.9040	1.0000	0.1460	0.6655	0.7159
47	聪明	寒冷	0.0150	0.6160	0.0120	0.2167	0.1404
48	红	粉红	0.8460	1.0000	0.0730	0.7079	0.8952
49	红	绿	0.7920	0.8611	0.1900	0.5606	0.6934
50	高兴	粉红	0.0920	0.6160	0.0000	0.2370	0.1719
51	高兴	开心	0.9620	1.0000	0.0380	0.9000	0.9795
52	年轻	老	0.7540	0.8611	0.1720	0.4295	0.6536
53	年轻	重	0.0520	0.6825	0.0510	0.3325	0.3943
54	厚	重	0.5620	1.0000	0.0470	0.3768	0.3794
55	炎热	干燥	0.5690	0.6825	0.1140	0.5329	0.5000
56	炎热	好	0.0380	0.6160	0.0170	0.2813	0.2149
57	初级	基础	0.8690	0.6825	0.1100	0.2932	0.3443
58	初级	高级	0.7230	0.7741	0.1350	0.3432	0.3813
59	冷	凉	0.8420	1.0000	0.0970	1.0000	1.0000
60	巨大	新	0.0310	0.6160	0.0750	0.3867	0.2883

注：S_{m}为人工打分的结果，S_1–S_4分别对应方法 1–方法 4 的结果

表 2 实验参数设置

公式调节参数		缩放系数		组合符号参数	
参数	取值	参数	取值	参数	取值
α	0.50	ω_{11}	1.20	λ_1	0.55
β	1.00	ω_{12}	1.10	λ_2	0.45
		ω_{21}	2.00	λ_3	0.40
		ω_{22}	1.70	λ_4	0.40
		ω_{23}	1.50	λ_5	0.50
		ω_{24}	1.20	λ_6	0.40
		ω_{31}	1.20	λ_7	0.60
		ω_{32}	1.05	λ_8	0.50
		ω_{41}	1.50	λ_9	0.60
		ω_{42}	1.20	λ_{10}	0.70
		ω_{43}	1.10	λ_{11}	0.80
				λ_{12}	0.60

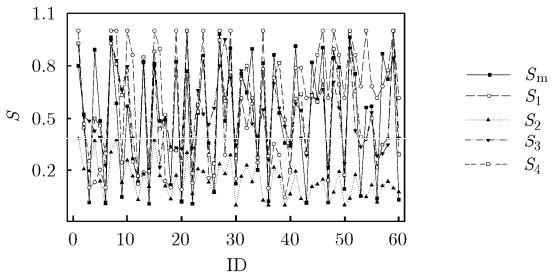

图 2 各方法相似度结果

表 3 词语相似度计算部分结果

ID	W_1	W_2	S_{m}	S_1	S_2	S_3	S_4
11	男人	母亲	0.2620	0.8611	0.1650	0.3879	0.2485
51	高兴	开心	0.9620	1.0000	0.0380	0.9000	0.9795
57	初级	基础	0.8690	0.6825	0.1100	0.2932	0.3443
58	初级	高级	0.7230	0.7741	0.1350	0.3432	0.3813

注：S_{m}为人工打分的结果，S_1–S_4分别对应方法 1–方法 4 的结果

表 3 则是从表 1 中选取的部分代表性结果。

从图 2 和结果表中可以看出，方法 2 的计算结果在数值上普遍偏低，与人工打分的符合情况不太理想，主要是因为基于上下文特征的方法考虑众多特征，加上一些噪声的引入，从而造成高维向量的

相似度普遍较低；另外，3 种方法其数值跨度较大(数据点在纵轴上分布范围广)，并且与人工打分均有一定符合，这是因为基于语义词典的方法往往在某些方面与人工思维符合程度较好。方法 1 的计算结果数值离散度较小，而另外 3 种方法的结果则很少重复，主要是因为方法 1 计算时所考虑的因素较少，计算公式中各项指标取值范围有限，而方法 2 基于高维向量计算，其结果分布较广，方法 3 和方法 4 则考虑了较多因素，能够有较好的区分度。方法 1，方法 3 和方法 4 在不同的词语对上表现有一定差别，而方法 3 和方法 4 总体表现比较接近。例如，“男人 母亲”这一对词语的计算结果上，方法 3 和方法 4 较为接近人工打分，而方法 1 有较大偏差，“初级 高级”这一对词语的情况则相反，而“高兴 开心”这一组词语 3 种方法的结果均比较接近，这也体现了《知网》和概念基元符号系统在设计理念上的异同。另外，同样基于概念基元符号系统，方法 4 的结果在总体上比方法 3 更贴近人工打分(图中方法 4 的线形及数据点分布趋势与人工打分更为贴近)，这说明本文综合考虑语义网络的各项特性以及考虑节点层次的权重是有效的。

总体上，本文的计算结果与人工打分比较符合，不过其中也存在有较大偏差的项目。例如，“初级”(gu30aac21)和“基础”(ru12eb1,j721)应当具有较高相似性，但其计算结果却相似性较低，主要是因为“初级”的概念内涵采用“30aac21”来表示，而“基础”的概念内涵则是“j721”和“12eb1”的组合，两者之间符号层面差异明显，且尚未建立概念关联关系。

对各方法的计算结果进行定量分析，这里考虑 3 个指标——兼容度(Compat)[16]、相关系数(r)[4]和序对符合度(Opc)。兼容度用于绝对符合程度考察，相关系数和序对符号度用于考察相对符合程度。序对符合度的计算公式为

$$\mathrm{Opc}=\sum_{(\mathrm{WP}_i,\mathrm{WP}_j)} E(\mathrm{WP}_i,\mathrm{WP}_j)\Big/C_N^2$$

其中，$E(\mathrm{WP}_i,\mathrm{WP}_j)$ 为序对一致(指排序对中两两相对位置一致)的示性函数表示(即序对一致时函数取值为 1，否则取 0)，分母 C_N^2 表示从 N 个元素取出 2 个的组合数，为两两序对总数。

本文使用阈值 $\theta=0.4$，计算各方法的兼容度、相关系数和序对符合度(方法 2 数值偏差较大，不计算兼容度)，其结果如表 4 所示。

由表 4 可知，所有方法的相关系数均大于 0.5，在统计上则认为它们与人工打分均具有中等强度以上的相关性，也反映出各方法与人工结果的符合情况较好。相比于方法 2，另外 3 种方法的相关系数明显更高，说明基于语义词典的方法与人工判断比较容易达成一致；不过在序对符合度上，其差距并没有相关系数上的明显，方法 1 与其表现相当，说明仅考虑排序情况，方法 2 也是可行的。方法 3 的相关系数与方法 1 相当，但兼容度明显好于方法 1，说明该方法在取值上与人工打分的总体偏差较小。本文方法在 3 个指标上均优于其他方法，这说明本文方法在计算结果上与人工判断有更好的符合程度，从而可知，基于概念基元符号系统进行词语相似度计算也是有效的。最后，本文还对计算结果与人工打分结果进行了相关性检验，在原假设“两组数据不相关”下，采用 Spearman 非参数检验计算得到 p 值为 8.65×10^{-13}，说明待检验数据是显著相关的，进一步验证了本文结果与人工判断的符合程度。

表 4 各方法的评价指标结果

算法	Compat	r	Opc
方法 1	0.645	0.618	0.679
方法 2	–	0.554	0.684
方法 3	0.721	0.637	0.730
方法 4	0.812	0.786	0.775

4.3 简单应用

词语相似度计算在自然语言处理、机器翻译等多个领域有重要作用。以基于实例的机器翻译为例，假设待翻译句子“律师开展的调查”，经过搜索，在实例库中找到两个翻译实例：

(1) 警察开展的调查 /the investigation conducted by the police。

(2) 去年开展的调查 /the investigation conducted last year。

经对比计算，“律师”和“警察”的相似度为 0.3775，和“去年”的相似度为 0.1769，故选用实例(1)进行类比翻译，从而得到正确的译文：the investigation conducted by the lawyer。

5 结束语

本文以概念基元符号系统为基础，提出一种基于语义词典的相似度计算方法，从该符号系统的设计理念出发，并充分挖掘其中各项信息，包括层次性、网络性、对比对偶特性、挂靠特性和五元组信息，最终形成一个多维度的计算公式；另外，为节点深度和节点距离赋予权重的做法使之与实际情况更加符合。采用本文提出的方法在人工构建的测试集上进行实验，并与其他方法进行比较，结果表明本文方法计算的相似度与人工打分符合情况最好，在定量评价指标上也取得了最优的结果，兼容度、相关系数和序对符合度分别达到 0.812, 0.786 和

0.775。

概念关联性在本文中只能通过概念关联式体现，而已构建的关联式规模尚小，许多潜在关联性并没有得到应用，因此，下一步需要继续挖掘和构建概念关联式，完善关联式集合；与此同时，也有必要尝试提出新的度量关联性的方法。另外，该方法目前只能对包含在词典内的词语进行处理，对于未登录词则无能为力，未来很有必要探索该方法下未登录词的处理，以扩大该方法的适用性，这也是进一步的工作内容。

参 考 文 献

[1] LIN D. An information-theoretic definition of similarity semantic distance in WordNet[C]. Proceedings of the 15th International Conference on Machine Learning, San Francisco, CA, USA, 1998: 296–304.

[2] WU Z and PALMER M. Verbs semantics and lexical selection [C]. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 1994: 133–138. doi: 10.3115/981732.981751.

[3] RESNIK P. Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language[J]. *Journal of Artificial Intelligence Research*, 1999, 11(7): 95–130. doi: 10.1613/jair.514.

[4] 王桐, 王磊, 吴吉义, 等. WordNet 中的综合概念语义相似度计算方法[J]. 北京邮电大学学报, 2013, 36(2): 98–101. doi: 10.13190/jbupt.201302.98.wangt.
WANG Tong, WANG Lei, WU Jiyi, *et al*. Semantic similarity calculation method of Comprehensive concept in WordNet[J]. *Journal of Beijing University of Posts and Telecommunications*, 2013, 36(2): 98–101. doi: 10.13190/jbupt.201302.98.wangt.

[5] WANG Junhua, ZUO Wanli, and PENG Tao. Hyponymy graph model for word semantic similarity measurement[J]. *Chinese Journal of Electronics*, 2015, 24(1): 96–101. doi: 10.1049/cje.2015.01.016.

[6] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[C]. 第三届汉语词汇语义学研讨会论文集, 台北, 中国, 2002: 59–76.
LIU Qun and LI Sujian. Words semantic similarity computation based on HowNet[C]. Proceedings of the 3rd Chinese Lexical Semantics Workshop, Taipei, China, 2002: 59–76.

[7] 李国佳. 基于知网的中文词语相似度计算[J]. 智能计算机与应用, 2015, 5(3): 49–52. doi: 10.3969/j.issn.2095-2163.2015.03.015.
LI Guojia. Chinese words similarity computation based on HowNet[J]. *Intelligent Computer and Applications*, 2015, 5(3): 49–52. doi: 10.3969/j.issn.2095-2163.2015.03.015.

[8] 张沪寅, 刘道波, 温春艳. 基于《知网》的词语语义相似度改进算法研究[J]. 计算机工程, 2015, 41(2): 151–156. doi: 10.3969/j.issn.1000-3428.2015.02.029.
ZHANG Huyin, LIU Daobo, and WEN Chunyan. Research on improved algorithm of word semantic similarity based on HowNet[J]. *Computer Engineering*, 2015, 41(2): 151–156. doi: 10.3969/j.issn.1000-3428.2015.02.029.

[9] 孙晶, 张东站. 基于逆概念频率的词语相似度计算[J]. 厦门大学学报(自然科学版), 2015, 54(2): 257–262. doi: 10.6043/j.issn.0438-0479.2015.02.018.
SUN Jing and ZHANG Dongzhan. Word similarity computing based on inverse concept frequencies[J]. *Journal of Xiamen University* (*Natural Science*), 2015, 54(2): 257–262. doi: 10.6043/j.issn.0438-0479.2015.02.018.

[10] BROWN P, PIETRA S, PIETRA V, *et al*. Word sense disambiguation using statistical methods[C]. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA, 1991: 264–270. doi: 10.3115/981344.981378.

[11] 关毅, 王晓龙. 基于统计的汉语词汇间语义相似度计算[C]. 第七届全国计算语言学联合学术会议论文集, 哈尔滨, 中国, 2003: 221–227.
GUAN Yi and WANG Xiaolong. A statistical measure of semantic similarity between Chinese words[C]. Proceedings of the 7th Joint Symposium on Computational Linguistics, Harbin, China, 2003: 221–227.

[12] 王石, 曹存根, 裴亚军, 等. 一种基于搭配的中文词汇语义相似度计算方法[J]. 中文信息学报, 2013, 27(1): 7–14. doi: 10.3969/j.issn.1003-0077.2013.01.002.
WANG Shi, CAO Cungen, PEI Yajun, *et al*. A collocation based method for semantic similarity measure for Chinese words[J]. *Journal of Chinese Information Processing*, 2013, 27(1): 7–14. doi: 10.3969/j.issn.1003-0077.2013.01.002.

[13] 李慧. 词语相似度算法研究综述[J]. 现代情报, 2015, 35(4): 172–177. doi: 10.3969/j.issn.1008-0821.2015.04.035.
LI Hui. A review on the research of word similarity algorithms[J]. *Journal of Modern Information*, 2015, 35(4): 172–177. doi: 10.3969/j.issn.1008-0821.2015.04.035.

[14] 黄曾阳. HNC 理论全书(第五册)[M]. 北京: 科学出版社, 2015: 1–102.
HUANG Zengyang. The Complete Book of Hierarchical Network of Concepts Theory (Book 5)[M]. Beijing: Science Press, 2015: 1–102.

[15] 苗传江. HNC(概念层次网络)理论导论[M]. 北京: 清华大学出版社, 2005: 1–49.
MIAO Chuanjiang. Introduction to HNC Theory[M]. Beijing: Tsinghua University Press, 2005: 1–49.

[16] 吴佐衍, 王宇. 基于 HNC 理论的词语相似度计算[J]. 中文信息学报, 2014, 28(2): 37–43. doi: 10.3969/j.issn.1003-0077.2014.02.005.
WU Zuoyan and WANG Yu. A new measure of semantic similarity based on hierarchical network of concepts[J]. *Journal of Chinese Information Processing*, 2014, 28(2): 37–43. doi: 10.3969/j.issn.1003-0077.2014.02.005.

[17] 史燕. 基于 HNC 的汉语句子相似度算法的研究[D]. [硕士论文], 江苏大学, 2009: 14–19. doi: 10.7666/d.y1604350.
SHI Yan. The research on Chinese sentence similarity algorithm based on HNC[D]. [Master dissertation], Jiangsu University, 2009: 14–19. doi: 10.7666/d.y1604350.

池哲洁： 男，1988 年生，博士，研究方向为自然语言处理.

张 全： 男，1968 年生，研究员，研究方向为自然语言理解、语言知识处理.